Selecting Research and Other Works for Deposit to the Stanford Digital Repository

When it is time to preserve the products of your research and creative output, you will need to carefully consider exactly which components of your work are most suitable for archiving and sharing. Different disciplines conduct research in different ways and produce content in different forms, so we have provided some general guidelines below that apply to most researchers, as well as specific examples for various fields of study, including science & engineering, social sciences, humanities, geospatial studies, and music composition. If you have questions about the general guidelines, please contact us at sdr-contact@lists.stanford.edu. For questions about specific areas of study, please contact the individual listed for the appropriate category below.

GENERAL GUIDELINES

- **Definitely deposit:**
  - original data sets, original software code, raw data obtained from analysis of physical samples, observational data that can not be regenerated
  - data sets that are not original but that are not easily available online and that you have permission to share
  - for social science data, include study descriptions, codebooks, and summary statistics

- **Maybe deposit:**
  - intermediate versions of analyses or code if they are potentially useful to others or were used in publications or theses

- **Not necessary to deposit:**
  - incomplete, non-functional, or intermediate versions of code that would be of marginal usefulness to others
  - output files from analyses if 1) the data set and code used to generate the output are deposited and 2) regenerating the output from the deposited files is fairly easy to do
  - data sets that are preserved and accessible via other institutions or organizations
  - graphs or charts created from the original data that could easily be regenerated

- **Do not deposit:**
  - any data that contains personal identifying information for human subjects

- **Exceptions:**
  - Output files from analyses may be deposited if they are time-intensive to regenerate or are not excessively large, or can not be easily recreated from the deposited data set and code.

SCIENCE & ENGINEERING RESEARCH

For Science & Engineering-specific questions, contact Amy Hodge, Science Data Librarian (amyhodge@stanford.edu)

Example 1: Your work involves writing software code to analyze a particular set of data. The data set used for the analysis is not easily available online.
• Definitely deposit:
  ○ data set used for the analysis
  ○ final version of the software code

• Maybe deposit:
  ○ additional versions of the software code that were used in publications or theses or that may be of use to other researchers

Example 2: Your work involves analysis of Kepler mission data.
• Definitely deposit:
  ○ analytical files and software code
  ○ metadata that identifies exactly which Kepler data were used for your analysis
• Not necessary to deposit:
  ○ Kepler mission data, as these are managed by NASA

Example 3: Your work involves the analysis of physical samples, using equipment that is difficult to gain access to, expensive to run, or causes destruction of the sample.
• Definitely deposit:
  ○ raw data generated from sample analysis, as re-collection of these data may be difficult, costly, or impossible
  ○ results from the sample analysis, including a description of the steps taken during the analysis to achieve the final results
  ○ software code or scripts written specifically for the analysis
• Maybe deposit:
  ○ other software code that may be difficult for others to find or acquire, as long as it is not proprietary, does not require a license to use, and you have permission to deposit and share it

Example 4: Your work involves the collection of observational data associated with a particular time and/or place.
• Definitely deposit:
  ○ original observational data
  ○ metadata adequate for others to understand the observations
• Maybe deposit:
  ○ cleaned up versions of the observational data
  ○ results of analyses that would be timely or costly to recreate, especially if the files are not excessively large or support a publication

Example 5: Your work involves biomedical studies with patient health information
• Definitely deposit:
  ○ anonymized or de-identified data used for the study
  ○ raw data used for the study ONLY if they do not contain identifying information
  ○ outputs from your analysis
• Do not deposit:
  ○ raw data if they contain identifying information (see section below on identifying information)
SOCIAL SCIENCES RESEARCH
For social science-specific questions, contact Ron Nakao, Data and Computational Social Science Librarian (ronbo@stanford.edu)

Example 6: Your work involves analyzing census data
- Definitely deposit:
  - study descriptions, codebooks, and summary statistics
  - results of your analysis
- Do not deposit:
  - original census data, as these are easily accessible elsewhere

Example 7: Your work involves assembling data from a variety of sources for analysis
- Definitely deposit:
  - study descriptions, codebooks, and summary statistics
  - descriptions of the sources of your data
  - de-identified or anonymized versions of raw data that you have permission to share
  - results of your analysis
- Do not deposit:
  - data from other sources that you do not have permission to share, requires a license to use, or are otherwise restricted
  - any data that contain identifying information

Example 8: Your work involves the collection and analysis of interviews
- Definitely deposit:
  - study descriptions, codebooks, and summary statistics
  - audio and/or transcripts of the interviews that have been edited and/or redacted in order to de-identify or anonymize the files
  - results of your analysis
- Do not deposit:
  - any data that contains identifying information

HUMANITIES PROJECTS
For humanities-specific questions, contact Hannah Frost, SDR Service Manager (hfrost@stanford.edu)

Example 9: You have created a visualization of slave transactions that are recorded in documents held in a national archive
- Definitely deposit:
  - compiled digital database used to create the visualization
  - information describing the contents of the database
- Maybe deposit:
  - code or web content used to create the visualization
- Do not deposit:
  - original documents held in the national archive

Example 10: Your work involves an analysis of a museum newspaper archive. You are not allowed to share the corpus of documents from the newspaper archive.
- Definitely deposit:
  - scripts you created for the analysis
  - results of your analysis
• Do not deposit:
  ○ corpus from the newspaper archives

Example 11: Your work involves an analysis of tax documentation. You are not allowed to share the tax documents.
  • Definitely deposit:
    ○ information about the research process
    ○ tools you created to perform the analysis
    ○ scripts you created to analyze the documents
  • Do not deposit:
    ○ original tax documentation

MUSICAL/PERFORMANCE WORKS
For musical and performance work-related questions, contact Jon Manton, Stanford Archive of Recorded Sound (jmanton@stanford.edu)

Example 12: Your work involves creation of a multichannel multimedia performance piece
  • Definitely deposit:
    ○ a mixed down capture of a performance, preferably the first performance, and any subsequent performances you might like
    ○ a video of a performance, if available
    ○ the final version of the software code/applications created for performance, including MAX/MSP patches, etc.
    ○ images of any custom instruments or performance devices created
  • Not necessary to deposit (but should be backed up elsewhere):
    ○ individual multi-track files
    ○ unedited video material
    ○ digital audio workstation (DAW) files from Nuendo, Pro-Tools, etc.

Example 13: Your work involves the creation of an instrumental or stage work
  • Definitely deposit:
    ○ copy of final score
    ○ copy of final libretto
    ○ program from first performance
    ○ audio/video recording of first performance
  • Maybe deposit:
    ○ composition drafts
    ○ libretto drafts
  • Not necessary to deposit:
    ○ short sketches of phrases etc.

Example 14: Your work involves a musicological study
  • Definitely deposit:
    ○ a copy of the final study
    ○ completed copies of any published articles written by you as part of your study or subsequently
    ○ audio/video interviews conducted as part of your study (with releases signed by interviewees); be sure you have permission from interviewees to share this information and/or that they have been properly de-identified or anonymized
● Not necessary to deposit:
  ○ existing published materials referenced or cited as part of your study

GEOSPATIAL DATA
For GIS-specific questions, contact Kim Durante, Metadata Analyst (kdurante@stanford.edu)

● All GIS layers should have a defined coordinate system (i.e. the "reference system description" in the layer properties should not be "undefined").
● Metadata for all tables containing latitudes and longitudes should include the DATUM.
● Files should include a data dictionary that defines data headings and numerical units.
● For data tables used as part of a feature class, pre-join the table to its geometry before depositing.
● Definitely deposit:
  ○ geospatial and attribute information that you collected
  ○ final version of the files with the geospatial and attribute information joined
● Maybe deposit:
  ○ geospatial information that you obtained from another source that is not readily available online or that is publicly-available but that Stanford may not be preserving elsewhere
  ○ attribute information that you did not collect, depending on the source
● Do not deposit:
  ○ base layers that are not directly part of the research data collection
  ○ base layers that are easily available from other sources, like USGS elevation data or basic administrative boundary data
  ○ data for which Stanford has a license, like ESRI Data & Maps layers

ABOUT IDENTIFYING INFORMATION, especially for Biomedical or Social Science research

● Direct identifiers should be removed or masked prior to depositing. These include: names, addresses (including zip codes), phone numbers (including area codes), social security numbers, drivers license numbers, certification numbers etc.
● Indirect identifiers, such as occupation, dates of significant events, job history, educational institutions, rare diseases, place of medical treatment or doctor giving care, and other types of information that could be used in conjunction with other information to identify individuals may need to be recoded in order to minimize the risk of disclosure.
● Respondent identifiers used for the study should also be removed before depositing data.
● If it is not possible to remove or recode all identifying information without significantly impacting the usability of the data, you may not be able to deposit or share the data.
● You should also be aware of and conform to any policies or procedures set forth by your local IRB (Institutional Review Board).
● Additional guidance is available in ICPSR's "Guide to Social Science Data Preparation and Archiving" (https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/index.html) and in the British Medical Journal article "Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers" (http://dx.doi.org/10.1136/bmj.c181). For questions on this topic relating to research at Stanford, please visit Stanford’s Research Compliance Office (http://humansubjects.stanford.edu/).